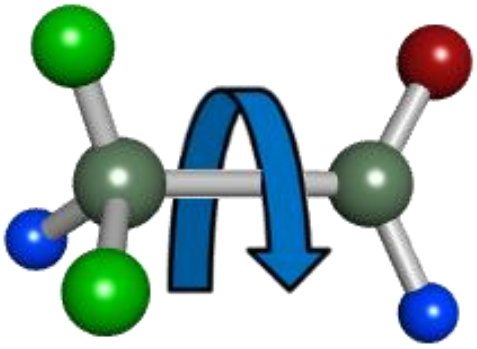
COSMO*logic* GmbH & Co. KG
Computational Chemistry and Fluid Thermodynamics

COSMO*conf*X User Guide

Version 3.0 (Jul 2013)



Copyright by
COSMO*logic* GmbH & Co. KG
Imbacher Weg 46, 51379 Leverkusen
Germany

cosmotherm@cosmologic.de
www.cosmologic.de

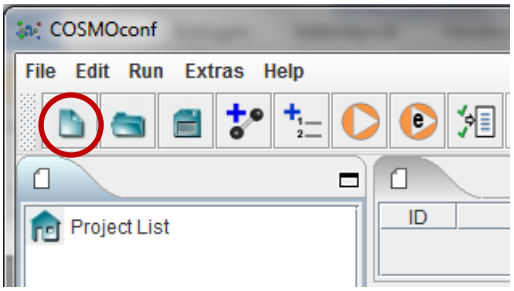
1 Quickstart 3
1.1 Introduction 3
2 Introduction to the COSMO*conf*X graphical user interface (GUI) 6
2.1 General 6
2.2 Projects and jobs 6
2.3 The Start Set and Results Set panel 8
2.4 The JOB DEFINITION panel 9
2.5 Running jobs locally or remote 11
2.6 Calculation time 13
2.7 Extracting results 13
2.8 The job status section 14
3 COSMO*conf* Linux command line version 14
3.1 Installation (Linux): 14
3.2 How to use the command line version? 14
3.3 Directories and file names 15
3.4 Example 16
4 Making your own job definition 16
4.1 The COSMO*conf* workflow 17
4.2 Unit operations / steps 18

1 Quickstart

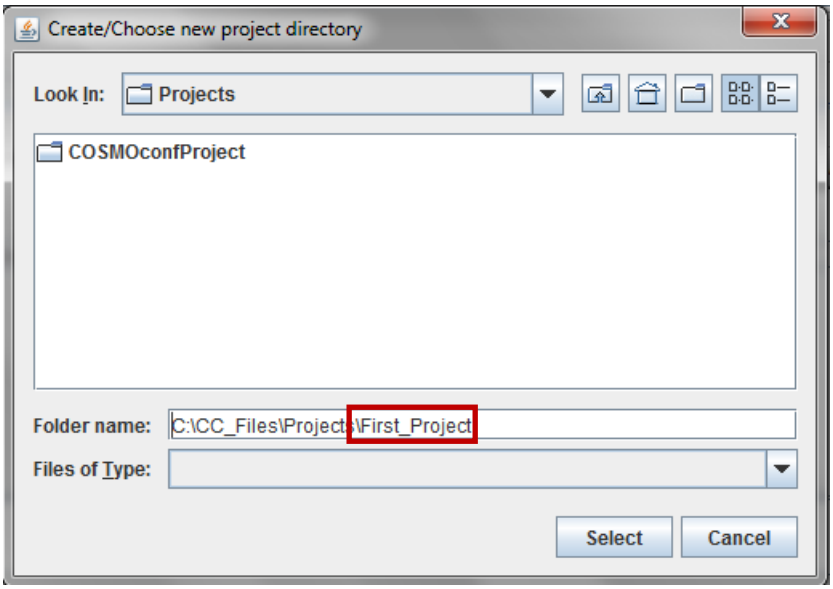
1.1 Introduction

This document will guide you through a standard COSMO*conf* calculation, i.e. a calculation using a pre-defined job template. In our example we will use the BP-TZVP-COSMO template for the creation of conformer COSMO files that can be used with the BP_TZVP_... parameterization of COSMO*therm*.

1) Create a project

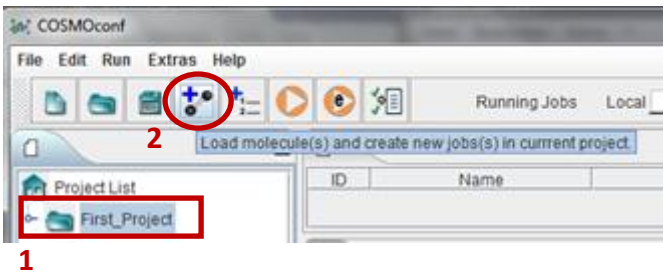


The project directory (existing directory) has to be selected in the file chooser. A new project directory (this is the name of our new project) can be added to the path. In the above example we create the project “First project”.

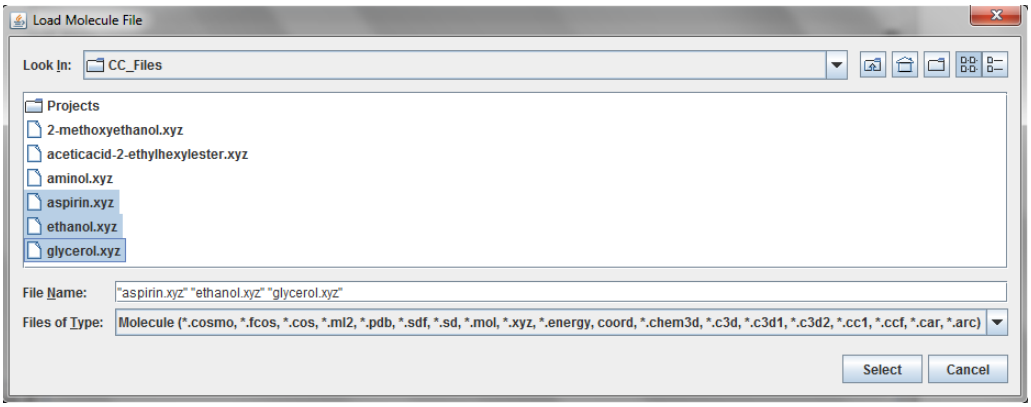


2) Load molecules

Select the new project (just click the project in the list) and click the “Load molecule(s)…” button.

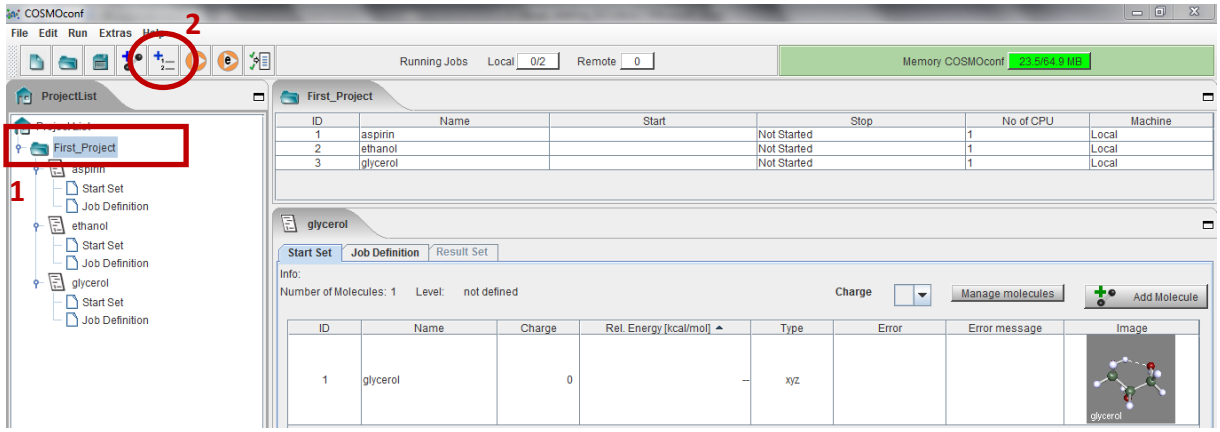


Now use the file chooser to load the structures of interest. In our case this is aspirin.xyz, ethanol.xyz, and glycerol.xyz. The reader accepts the structure file types of the “Files of Type” list. For file types that allow for 2D and 3D structures it’s important to use the 3D variant.

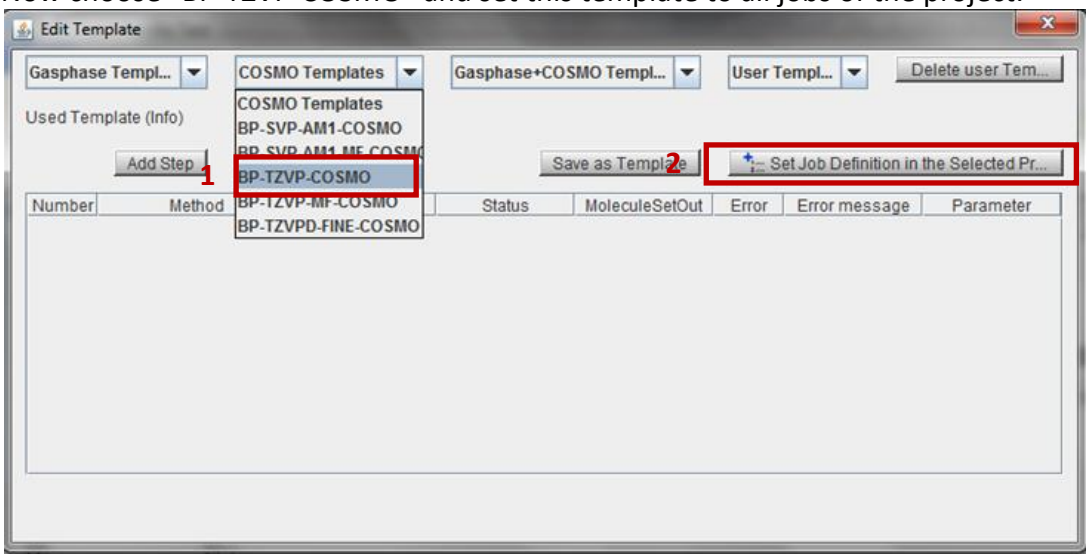


3) Set the JOB DEFINITION

In order to set the same JOB DEFINITION (this is the procedure for the conformer search) to all molecules. Select the project and use the “SET JOB DEFINITION…” button.

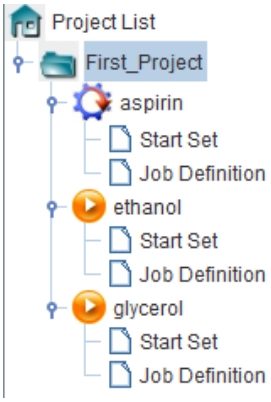


Now choose “BP-TZVP-COSMO” and set this template to all jobs of the project.



4) Run the project
Start the job on the local machine

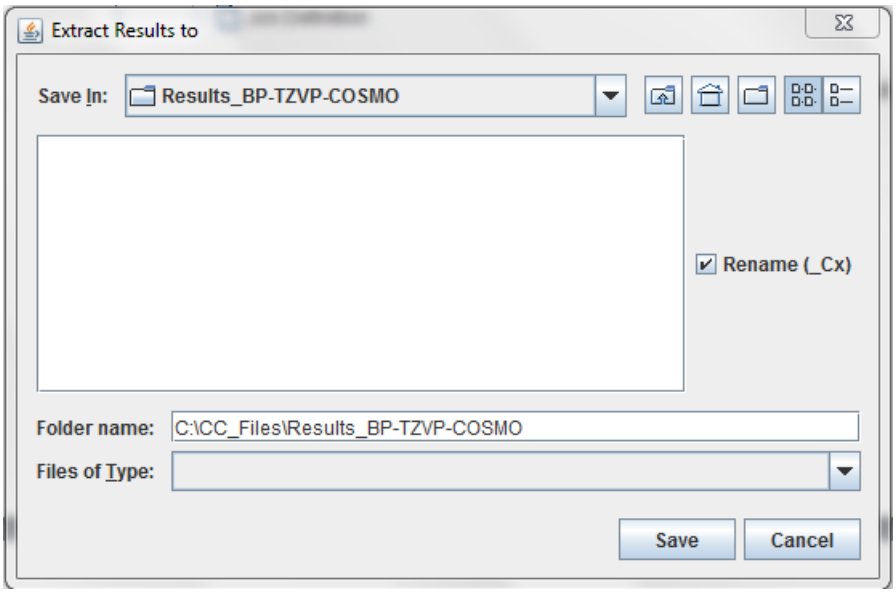
Because we allow for the use of one processor (default) the first job/molecule is running while the others are queued for execution.



Now we wait until all calculations have been finished (the icons should change to ✔).

5) Extract the results
Select the project and click the “Extract all results…” button

Now chose the result directory.



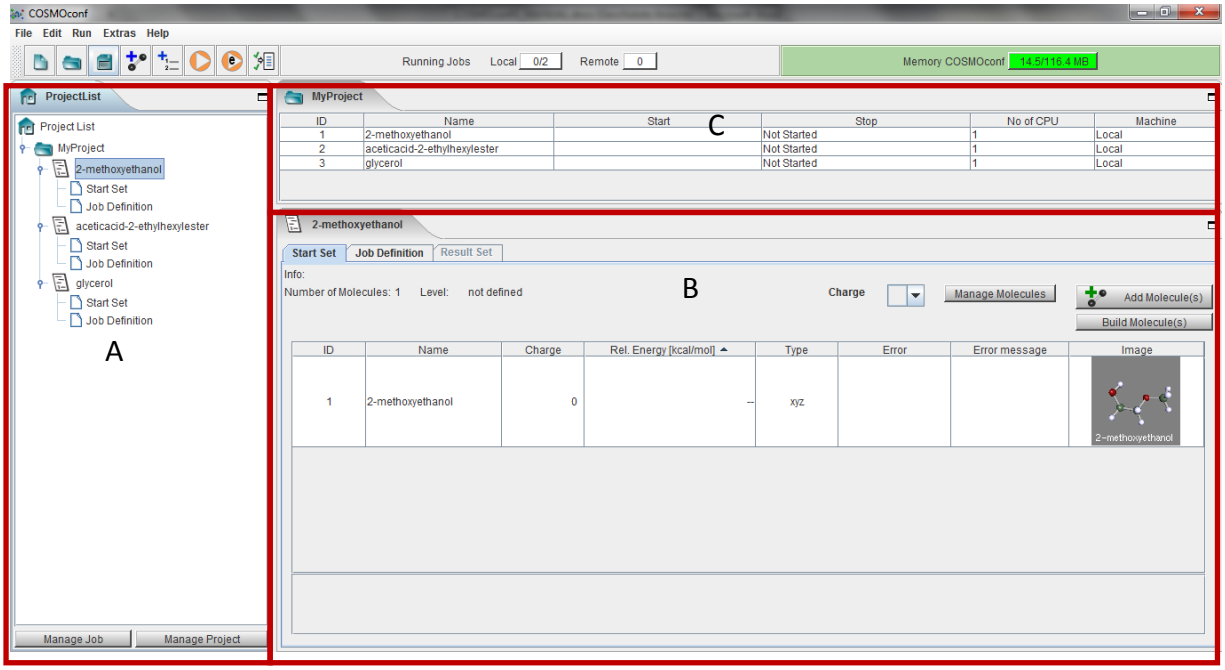
In our example we use a directory name that reflects the level of the conformers that have been created (“Results_BP-TZVP-COSMO”). The cosmo files of the conformers can be found in the chosen directory. By default the numbering follows the COSMO*therm* convention.

2 Introduction to the COSMO*conf*X graphical user interface (GUI)

2.1 General

The COSMO*conf*X GUI consists of three sections:

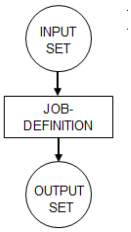
A) The project organization section
B) The job window contains the JOB DEFINITION, the START SET and one or more RESULT SETS.
C) The job status section



2.2 Projects and jobs

The project organization section holds the list of projects and their jobs. Each project might consist of one or more jobs of different types. In a standard COSMO*conf* calculation each job represents one molecules of interest.

A job can consist of one or more molecules (START SET) and a JOB DEFINITION. The JOB DEFINITION provides the information about the consecutive steps which will finally convert the input molecule(s) to the RESULT SET. Therefore, the user needs to define an input set and a job definition for every job. In most cases only one input structure, e.g. ethanol, is needed and COSMO*conf* will create a whole set of conformers.



Within a project the same JOB DEFINITION can be easily applied to all jobs or each job can have a different JOB DEFINITION. In the same way it is possible to submit all jobs of a project with a single click or to submit each job separately.

The tool bar on top of the GUI is a collection of actions that are necessary for a typical COSMO*conf* run. **All actions in this bar will work simultaneously on the whole project (all jobs of the project)**.

CREATE NEW PROJECT. The path to the project directory has to be chosen. If a project name is added to the path a subdirectory with the chosen name will be created. E.g.: we chose the “CC_Files” directory and add the project name ”First_Project”.

Open a closed project (open the *.cconf project file).

Save selected project.

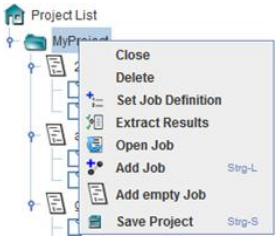
Create a new job in the actual project. Molecular 3D structure files can be picked with a file chooser menu. Each selected structure will create a new job in the selected project. The name of the structure file will be automatically used as job name. This name can be changed within the right mouse button menu (“Rename this calculation”).

Set job definition to the current project. All jobs of the project receive the same job definition (except the ones that have been defined before).

Save and run the project on the local computer

Save and run the project on the external computer (via network)

Extract all result files of the current project.



The context menu (right mouse button or MANAGE PROJECT button) can be used to close or delete the whole project. Use CLOSE if you like to be able to open the project again. The DELETE option will delete all data on the disk, i.e. the project *.cconf file, and the data of all corresponding jobs.

In addition the FILE menu offers the following options:

OPEN JOB: To open a previously closed job (choose the jobdefinition.xml file of the job).

ADD EMPTY JOB: Allows for the creation of an empty job with now molecules inside. The molecules can be added with the “Add” and “Build” buttons of section B later on.

In the typical COSMO*conf* usage (generation of a conformer set from one start structure) each molecule will be represented by a job. A job can be modified using the job context menu (right mouse button or the MANAGE JOB button). The available options depend on the status of the job:

VIEW JOB DIRECTORY	Open the job directory (the directory of the job execution) in a browser window. This option is mainly used for debug purposes.
RENAME THIS JOB	Change the name of the job. The name of the structure in the Input/Results Sets will not be changed, but the new name will be used in the result extraction (the extracted files will be named after the job).
CLOSE	The job will disappear from the list, but the data will be kept on disk. It can be re-opened later.
DELETE	All data will be deleted
REMOVE JOB FROM QUEUE	A queued job can be removed from the queue
STOP THIS JOB	The selected job will be stopped. It can be continued later.
EXTRACT RESULTS	The results of the selected jobs will be extracted to the user defined directory (see the “Extract results” paragraph below)
SAVE AND RUN THIS JOB	The current settings will be saved and the job will be started or queued. (see the “How to start a job?” paragraph below)

The following icons indicate the job status:

Job terminated properly

Erroneous job execution. See error message in “Job Definition” and/or “Molecule Set”

Job before execution

Job is running

Job in queue (waiting for a free processor).

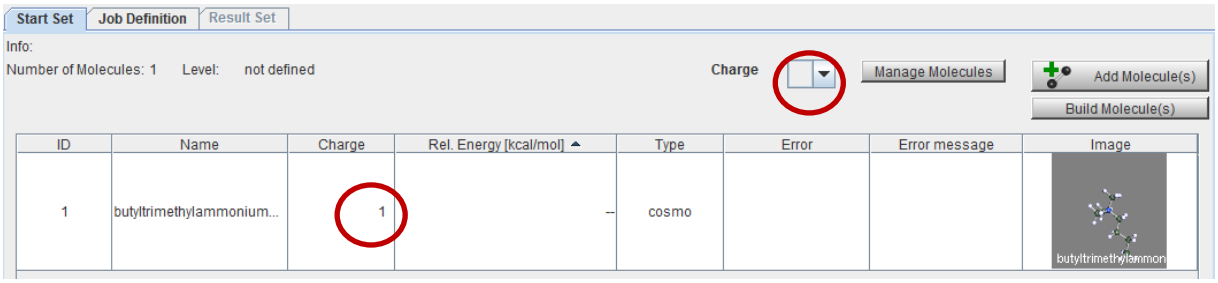
The job has been stopped by the user

Data transfer (from remote to local machine)

2.3 The Start Set and Results Set panel

The table can be sorted with respect to the REL. ENERGY (relative energy) by clicking on the column headline. Errors that occurred during the calculations are indicated by a negative number in the ERROR column and some information about the problem can be found in the “ERROR MESSAGE” column.

For start sets the molecular charge can be set in the CHARGE select box. The charge can also be set in the charge column of every single structure of the set.



A context menu (right mouse button or the MANAGE MOLECULES button) can be used to apply several options to one or several selected structures. Depending on the type of set (Start or Result) the option may vary.

SHOW MOLECULE	The structures of the actual selection will be displayed in a molecule viewer.
CREATE NEW CALCULATION	The selected structures will be used as "Start Set" for a new job. The job will be named after the first structure of the selection. This option works for all sets and can be used for subsequent treatment.
DELETE	Delete the selected structure (Only for jobs that have not been started yet).
EDIT MOLECULE	Start a structure editor to modify the current structure. (Only for jobs that have not been started yet)

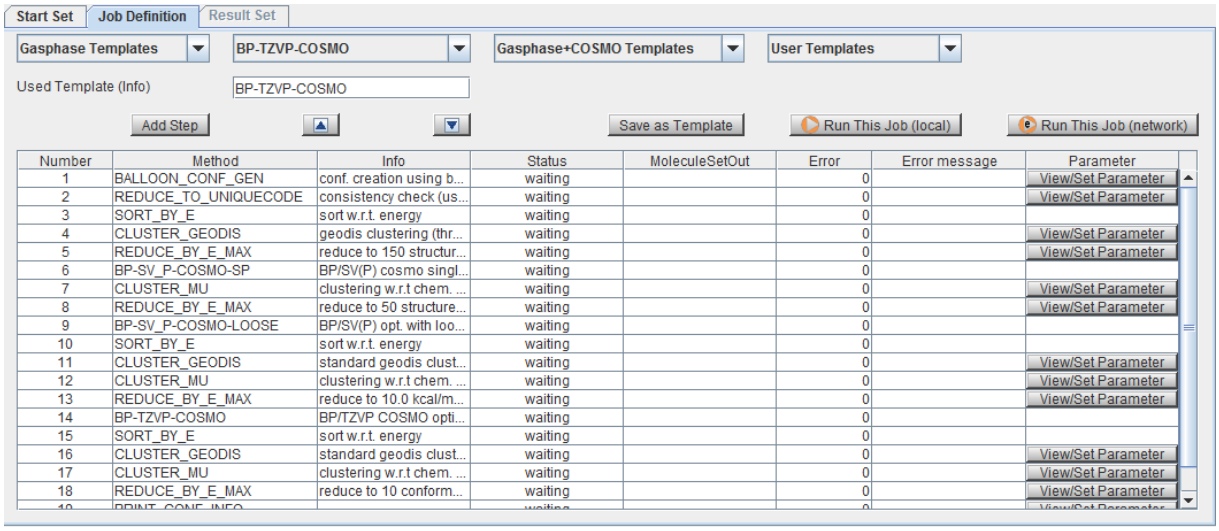
The

button of the START SET panel is similar to the "LOAD MOLECULE" button of the tool bar. The important difference is that it adds molecules to the current START SET instead of creating a new job. This option can be useful in a step by step procedure e.g. if one likes to perform a QM calculation on a set of input structures. To add more than one structure, however, is not useful for the standard conformer generation, because the current conformer generators are designed for one input structure only.

2.4 The JOB DEFINITION panel

Each job needs a job definition, i.e. a list of methods that should be applied successively. Because the job definition holds information about the status of the execution, it cannot be set or changed for running or terminated jobs. The job definition can be set for each job individually (context menu of the job or MANAGE JOB button) or for the whole project, depending on whether the project or the job is selected.

The status of jobs running on the local machine will be updated frequently. Remote jobs status information will not be updated during the calculation. The results will be copied at the end of the calculation and therefore the status will be updated only once at the end. The number of molecules (nmol=…) at the end of the step can be found in the status column.



Number	Method	Info	Status	MoleculeSetOut	Error	Error message	Parameter
1	BALLOON_CONF_GEN	conf. creation using b...	waiting		0		View/Set Parameter
2	REDUCE_TO_UNIQUECODE	consistency check (us...	waiting		0		View/Set Parameter
3	SORT_BY_E	sort w.r.t. energy	waiting		0		
4	CLUSTER_GEODIS	geodis clustering (thr...	waiting		0		View/Set Parameter
5	REDUCE_BY_E_MAX	reduce to 150 structur...	waiting		0		View/Set Parameter
6	BP-SV_P-COSMO-SP	BP/SV(P) cosmo singl...	waiting		0		
7	CLUSTER_MU	clustering w.r.t chem. ...	waiting		0		View/Set Parameter
8	REDUCE_BY_E_MAX	reduce to 50 structure...	waiting		0		View/Set Parameter
9	BP-SV_P-COSMO-LOOSE	BP/SV(P) opt. with loo...	waiting		0		
10	SORT_BY_E	sort w.r.t. energy	waiting		0		
11	CLUSTER_GEODIS	standard geodis clust...	waiting		0		View/Set Parameter
12	CLUSTER_MU	clustering w.r.t chem. ...	waiting		0		View/Set Parameter
13	REDUCE_BY_E_MAX	reduce to 10.0 kcal/m...	waiting		0		View/Set Parameter
14	BP-TZVP-COSMO	BP/TZVP COSMO opti...	waiting		0		
15	SORT_BY_E	sort w.r.t. energy	waiting		0		
16	CLUSTER_GEODIS	standard geodis clust...	waiting		0		View/Set Parameter
17	CLUSTER_MU	clustering w.r.t chem. ...	waiting		0		View/Set Parameter
18	REDUCE_BY_E_MAX	reduce to 10 conform...	waiting		0		View/Set Parameter

Default Templates

The default templates are divided into three groups: gasphase templates, COSMO templates and gasphase + COSMO templates. The gasphase templates will produce conformations as calculated in an ideal gas (i.e. vacuum) and the final results are `.energy` files. The COSMO templates will generate conformers relevant for the liquid phase and the final results will be `.cosmo` files.

If `.cosmo` and `.energy` files are needed we recommend using the combined templates. The calculation will be much faster than doing both calculations separately and the µ-clustering, which cannot be used in pure gasphase procedures, yields a conformer set especially adjusted for COSMO*therm* calculations.

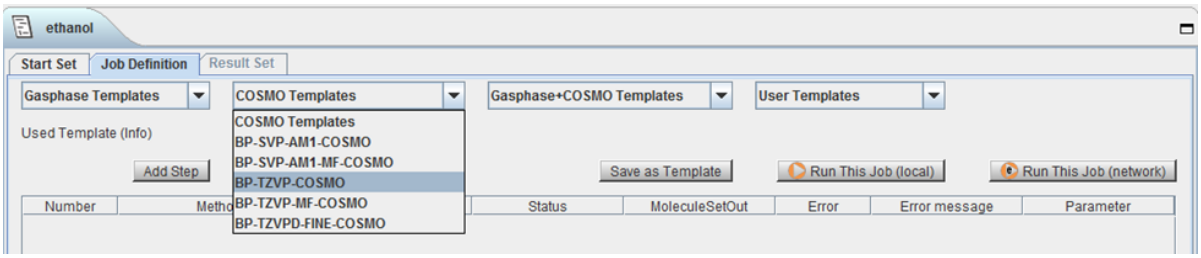
Within each group, the templates are further divided according to the details of the calculation.

- BP-SVP-AM1 indicates a quick semi-empirical AM1 geometry optimization with a BP-SVP single point density functional calculation. This level is very fast at the cost of some accuracy.
- BP-TZVP indicates a full geometry optimization with density functional theory and a medium sized basis set. This is the standard level for COSMO*therm*.
- BP-TZVPD-FINE indicates a full geometry optimization with density functional theory on BP-TZVP level, with a consecutive BP-def2-TZVPD single point. A FINE cavity is used in the COSMO calculations. The basis set is significantly larger and includes diffuse functions. This level is required for the COSMO*therm* BP-TZVPD-FINE parameterizations. **Note:** The BP-TZVP results are automatically generated on the fly during the calculation. The results will thus contains two full sets.
- MF marks the templates that make use of COSMO*frag* and MOPAC for the initial conformer generation. These are included for smaller molecules and compatibility to

previous versions. All templates not containing a MF will use BALLOON as a conformer generator.

Job Definition

There are two ways to define the procedure that should be used. The easiest and recommended way is the use of a predefined procedure, which can be chosen from the DEFAULT TEMPLATES lists.



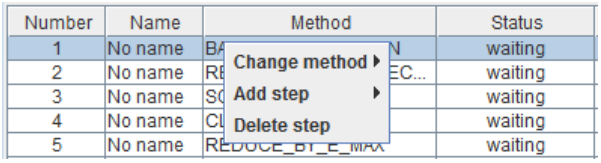
A more advanced option is the individual set up of the procedure. The GUI allows for user defined job definitions which can be stored as USER TEMPLATES with the SAVE AS TEMPLATE button. More details on modifying templates can be found in the "making your own job definition" section of this manual.

The jobs steps can be added with ADD STEP button which offers steps of different types.

The position of a step in the job can be changed with the blue arrows .

The parameters (if existent) of a job step can be viewed / changed by using the VIEW/SET PARAMETER button

Already defined steps can be changed via the right mouse button menu. This can also be used to modify the predefined default procedures.

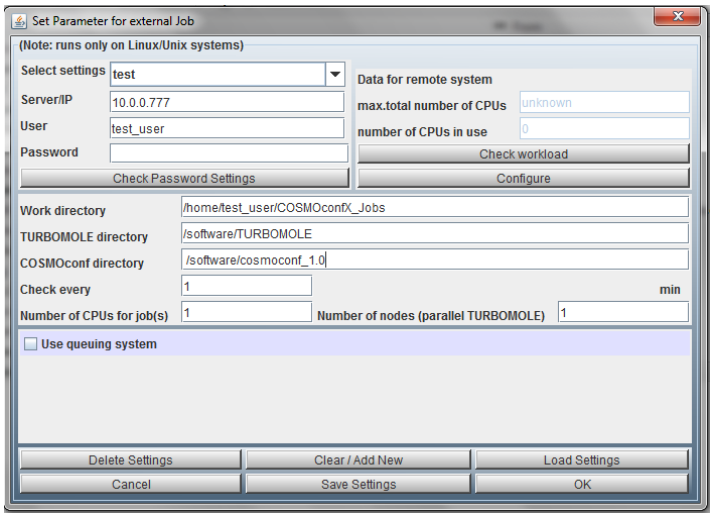


The user defined procedures can be saved (SAVE AS TEMPLATE) and used as USER TEMPLATES afterwards. Another way to define a new job template is the use of the "EDIT -> EDIT TEMPLATE" option of the menu bar. Here you can find an option for the deletion of a user defined job. Because the job definition is used to store the status of the single steps it cannot be changed for running, or finished jobs.

2.5 Running jobs locally or remote

Local jobs (jobs on the local machine) can be started without further settings. The only parameter that can be changed is the number of CPUs that should be used (see EXTRAS -> SETTINGS). This number defines the maximal number of jobs running at the same time. Surplus jobs will be queued. Because the work environment of a remote Linux system cannot be

known by the GUI, the user needs to define the settings. The settings for a machine can be saved (SAVE SETTINGS at the bottom of the menu) and re-used (SELECT SETTINGS). In order to check the password and login we recommend using the CHECK PASSWORD SETTINGS option.



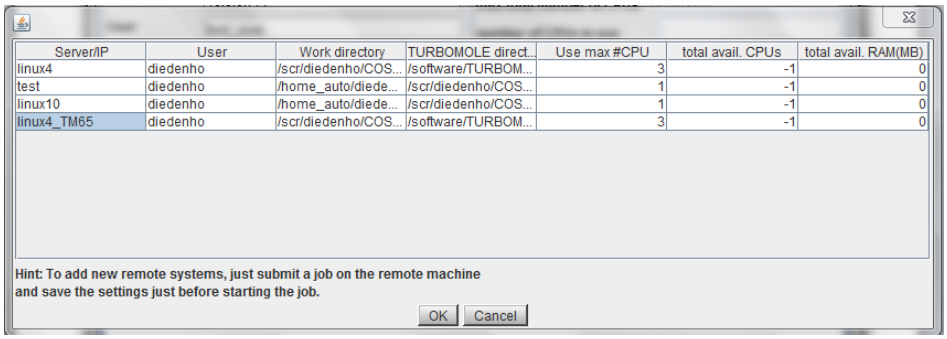
The paths that need to be adjusted are:

WORK DIRECTORY	Already existing directory that will be used for the COSMO*conf* calculations. **Please note** that the user needs read and write permissions in this directory.
TURBOMOLE DIRECTORY	Path to the TURBOMOLE installation. This path is named $TURBODIR in the TURBOMOLE documentation.
COSMOCONF DIRECTORY	Path to COSMO*conf* installation (beside other files this directory contains the cosmoconf_job_wrap.pl and the install script).

On the remote machine the jobs run like local jobs. The input data is transferred to the remote machine and the results are copied back to the local machine. Especially for bigger molecules this may take a few seconds (the transfer icon will appear). Therefore, the status of the job (see job definition panel) cannot be updated like for local calculations. The GUI does check the remote jobs at regular intervals, which can be defined by CHECK EVERY parameter. The maximum number of CPUs on the remote (i.e. the number of processes the GUI is allowed to send to the remote machine) is defined by NUMBER OF CPUS FOR JOB(S). The NUMBER OF NODES (parallel TURBOMOLE)" parameter determines the number of nodes that should be used for the parallel run of TURBOMOLE (for each job). Because each parallel execution generates communication overhead, we recommend running several serial jobs at the same time instead using the same amount of nodes for one parallel job. Please note: the "FINE level COSMO" calculations have not been parallelized yet. All job definitions containing the FINE level will be started as serial jobs automatically.

The current workload of the remote machine can be checked with the CHECK WORKLOAD option. The CONFIGURE button opens an overview of all defined remote machines. The TOTAL

AVAIL. CPUs can be used to ensure that the USE MAX # CPU value does not exceed the physical limits of the machine. This check is switched off "-1" by default. Remote machines can be deleted via the context menu.


Server/IP
User
Work directory
TURBOMOLE direct...
Use max #CPU
total avail. CPUs
total avail. RAM(MB)
linux4 diedenho /scr/diedenho/COS... /software/TURBOM... 3 -1 0
test diedenho /home_auto/diede... /scr/diedenho/COS... 1 -1 0
linux10 diedenho /home_auto/diede... /scr/diedenho/COS... 1 -1 0
linux4_TM65 diedenho /scr/diedenho/COS... /software/TURBOM... 3 -1 0
Hint: To add new remote systems, just submit a job on the remote machine
and save the settings just before starting the job.
OK
Cancel

2.6 Calculation time

COSMO*conf* uses quantum chemistry calculations for accurate results. Though density functionally theory is clearly a fast quantum chemistry method, the calculation of hundreds of geometry optimization may take quite some time. The following table provides a rough guideline on typical calculation times on a standard CPU.

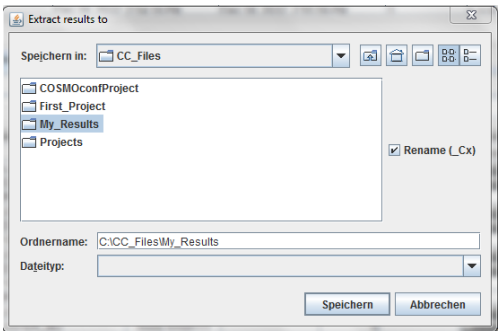
Number of atoms	Timescale
12	Minutes
20	Hours
40	Days
100+	Weeks

2.7 Extracting results

The result extraction method can be called for:

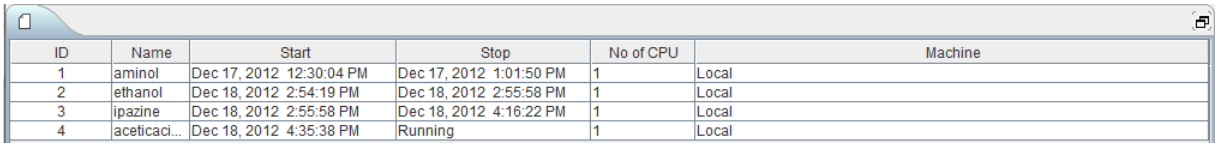
- Each job (context menu or MANAGE JOB button) individually.
- A whole project (context menu, MANAGE JOB button or the extract results icon from the tool bar), the results of all jobs of the selected project will be extracted.

The cosmo / energy files will be extracted to the chosen directory. By default the COSMO*therm* conformer nomenclature will be used (RENAME (…CX) activated). All molecules of an output set will be treated as conformers and numbered in the order of ascending energies. The name of the job will be used as base name of the conformer files. E.g.: the cosmo conformers of an "aspirin" job will be sorted and renamed to `aspirin_c0.cosmo`, `aspirin_c1.cosmo`, etc. The same holds for energy files. If the rename option is switched off the files will be copied without renaming i.e. the names of the output sets will be used.



2.8 The job status section

This panel gives a synopsis of the running and finished jobs of a project. In our example we have three finished and one running jobs, all on the local machine.



ID	Name	Start	Stop	No of CPU	Machine
1	aminol	Dec 17, 2012 12:30:04 PM	Dec 17, 2012 1:01:50 PM	1	Local
2	ethanol	Dec 18, 2012 2:54:19 PM	Dec 18, 2012 2:55:58 PM	1	Local
3	ipazine	Dec 18, 2012 2:55:58 PM	Dec 18, 2012 4:16:22 PM	1	Local
4	aceticaci...	Dec 18, 2012 4:35:38 PM	Running	1	Local

3 COSMO*conf* Linux command line version

All features of COSMO*conf* can be used from the command line to enables full batch processing capabilities. In addition a command line installation on a Linux computer is necessary to submit remote calculation from the GUI.

3.1 Installation (Linux):

A TURBOMOLE installation, version 6.4 or higher, is required for COSMO*conf* to work correctly.

Installation:

To ensure correct read, write and execute setting, the installation should be done by a member of the user group that will use the script later on.

1. Unpack the COSMO*conf* archive into a chosen directory

```
gunzip COSMOconf_....tar.gz

tar –xvf COSMOconf_....tar
```

2. Copy the license file (`license.ctd`) to the installation directory (the directory that has been chosen in step 1).

3. Change into the installation directory and start the COSMO*conf* installation script and follow the instructions.

```
./install
```

If the command line COSMO*conf* version should be used it might be convenient to include the COSMO*conf* directory (the one where you executed install) in the PATH. We recommend to define the new PATH in the local environment of the user (.bashrc,.cshrc etc.). For a bash user the entry looks like:

```
export PATH=<path to COSMOconf>:$PATH
```

3.2 How to use the command line version?

In order to do a series of calculations one needs to provide a directory with 3D input structures. The script needs a list of the structure input files, e.g.:

```
water.xyz

methanol.xyz

H3O+.xyz       +1
```

...

The molecular charge has to be given for charged molecules.
The script can be started as follows:

```
COSMOconf.pl -l <input list> -m <method> [-din <input dir.>] > <logfile>
```

Parameters in brackets are optional.

`<method >`	Describes the quantum mechanical level. A brief description can be found in the `cosmoconf.pl` help message. (simply execute cosmoconf.pl without arguments)
`<input dir.>`	Complete path of input file directory

List of allowed input file types

car	Accelrys/MSI Biosym/Insight II CAR format
cosmo	COSMO*logic* COSMO file
arc	MOPAC cartesian arc file
ml2	Sybyl Mol2 format
mol2	Sybyl Mol2 format
pdb	Unimolecular protein data bank format file
xyz	XYZ cartesian coordinates format
energy	COSMOlogic energy file
sdf	MDL Isis unimolecular 3D SDF V2000

3.3 Directories and file names

A calculation creates the following directories:

CMcal
This directory holds the subdirectories of the molecules, which contain all MOPAC[1], COSMO*frag*, and TURBOMOLE[2] calculations.

Results_of_job_...
These directories hold the final *.cosmo and *.energy files, respectively. The different conformers are numbered (_c0…_cn) accordingly to the COSMO data base convention. The conformers are ordered with respect to increasing energy. The file glucose_c0.cosmo, for instance, corresponds to the energetically (DFT energies) favorable conformer. **Please note**: the gas phase energies (*.energy files) have similar names, but the order corresponds to the

[1] MOPAC7 is the public domain version of:
MOPAC - A GENERAL MOLECULAR ORBITAL PACKAGE, ORIGINAL VERSION WRITTEN IN 1983 BY JAMES J. P. STEWART AT THE UNIVERSITY OF TEXAS AT AUSTIN, AUSTIN, TEXAS, MODIFIED TO DO ESP CALCULATIONS BY BRENT H. BESLER AND K. M. MERZ JR. 1989 locally modified by Andreas Klamt, COSMOlogic. For more details about MOPAC7, please visit http://sourceforge.net/projects/mopac7/

[2] TURBOMOLE, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989-2007, TURBOMOLE GmbH, since 2007; http://www.turbomole.com/

gas phase energies. Therefore, the gas phase structure of conformer name_c0.energy does not necessarily correspond to the COSMO conformer structure name_c0.cosmo.

Restart
The calculations can be restarted by using the same command in the same start directory. COSMO*conf* examines the already existent files and decides what to do.

3.4 Example

The following scheme explains the creation of COSMO files on the BP-TZVP-COSMO level:

1) Create 3D input structures, e.g. XYZ files.

2) Create a directory and copy the 3D files into this directory e.g.:

```
mkdir new_calc
cd new_calc
```

copy the files to `new_calc`

3) Create a list of the input file names (the file is called list hereafter).

Content of the file `list`:

```
ethanol.xyz
methanol.xyz
water.xyz
…
```

4) Start the script:

```
cosmoconf.pl –l list –m BP-TZVP-COSMO >list.log
```

The output of the script can be found in the file `list.log`. The COSMO files are collected in the `Results_of_job_BP-TZVP-COSMO` directory.

4 Making your own job definition

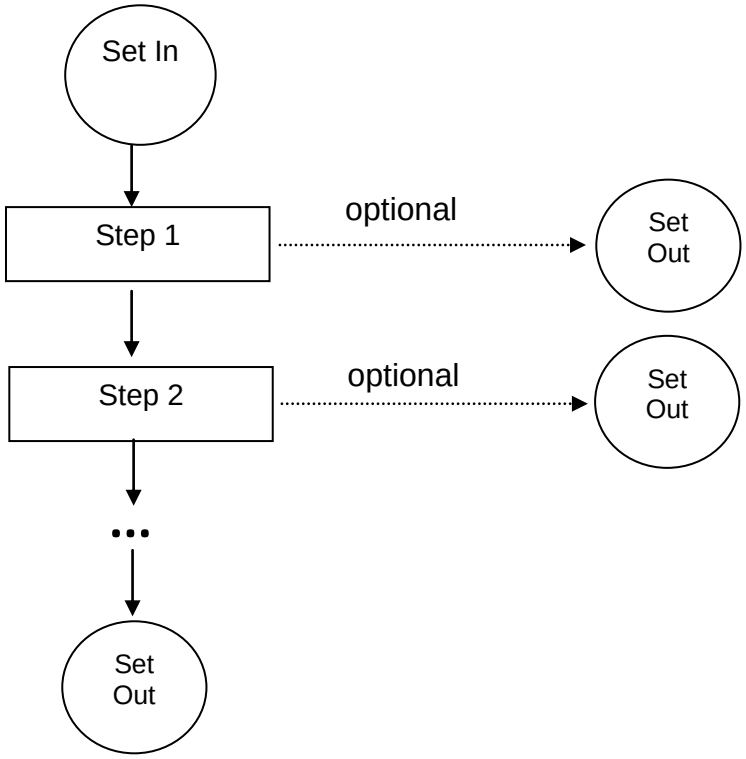
COSMO*conf* features a fully configurable calculation workflow to enable user defined calculation schemes. To efficiently use of these features some knowledge about xml and the different quantum chemistry levels as well as a fundamental understanding on conformer generation is recommended.

The default templates are constructed to yield good results for the majority of tasks, i.e. for organic compounds of small to medium size (1 to 60 Atoms). COSMO*conf* will work fine for larger molecules, but a user defined workflow might provide some benefits, either in calculation time or quality.

Some of the presented features can be accessed via the graphical user interface, other are only available from the command line.

4.1 The COSMO*conf* workflow

The COSMO*conf* workflow consists of unit operations (steps) working on sets of structures. The In/Out sets for these steps are lists of molecules / conformers in XML format. The results of the n^{th} step will be used as input for the $n+1^{th}$ step. Optionally intermediate molecule sets can be saved (this option currently not available from the GUI).



A typical workflow (and all default templates for JOB DEFINITIONS) will start with only a single structure and conduct the following basic steps:

1. Conformer generation, which can be either done by COSMO*frag* and MOPAC or by Balloon[3]. This step generates as many different structures as possible.
2. Check and Reduction: Throw out identical conformers, higher energy conformers, conformers with wrong stereochemistry and so on. There are multiple possibilities to select for those conformations needed.
3. Quantum Chemistry calculations: A single point or geometry optimization to provide information for better reduction or clustering.
4. Clustering: To select only those conformations that show a different physical behavior the SMS or µ-clustering routines can be used

The steps 2 to 4 are usually repeated several times with different settings to finally produce a small set of relevant conformers.

Apart from the above typical approach the user can define anything he needs. One possible example would be to use a whole set of conformations as a starting point, leave out the con-

[3] http://users.abo.fi/mivainio/balloon/. Mikko J. Vainio and Mark S. Johnson (2007) *Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm.* Journal of Chemical Information and Modeling, **47**, 2462 - 2474.

former generation with COSMO*frag* or Balloon and just do some reduction or clustering or qauntum chemistry

4.2 Unit operations / steps

The necessary information has been contracted in several tables:

Table 1: Lists the allowed steps. These are methods tags inside a step

Table 1a: Lists the allowed options and parameters for all steps of table 1.

Table2: General tags used outside steps for clean up or results extraction.

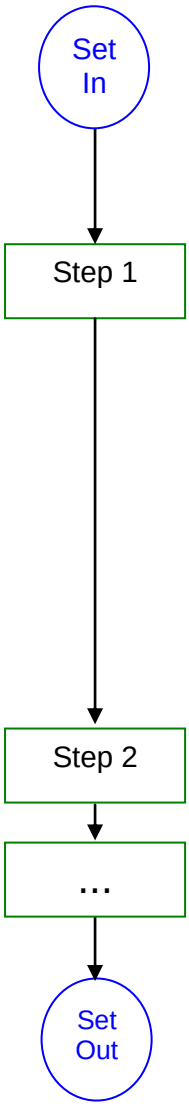
Tbale 3: Limitations of certain methods (e.g. conformer generator will work only on one structure)

Some steps allow for the definition of calculation type specific parameters. These options can be given in an extra tag (subtag of step).

Example: How to use parameter tags

```
<step>
        …
   <METHOD>PUT THE METHOD HERE</METHOD>
   <PARAMETER TAG>PUT THE PARAMETERS HERE</PARAMETER TAG>
  …
</step>
```

Example: Job definition XML




```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- general remarks:
   error number 0  -> normal termination
   error number <0 -> error, the error message should contain some
   description of the problem
-->
<job>
   <error>
      <number>0</number>
      <message></message>
   </error>
   <clean_up>1</clean_up>
   <info>first step (conf creation)</info>
   <molecule_set_in>cc_cluster_in.xml</molecule_set_in>
   <molecule_set_out>cc_cluster_out.xml</molecule_set_out>
   <job_schedule>
      <!--will be executed according to the step number-->
      <step>
      <!-- steps will be executed according to the step number -->
         <number>1</number>
         <!-- might be used in the output/error messages -->
         <info>conf. creation</info>
         <!-- file of output molecule set, not needed for input -->
         <molecule_set_out>step1_out.xml</molecule_set_out>
         <!-- method string of calculation -->
         <method>CF_MOPAC_CONF_GEN</method>
         <!-- status:waiting|running|ready -->
         <status>waiting</status>
         <error>
            <number>0</number>
            <message></message>
         </error>
      </step>
      <step>
         <number>2</number>
         <!-- just a name used in the output/error messages -->
         <info>cluster. creation</info>
         <!-- just a name used in the output/error messages -->
         <!-- file of output molecule set, not needed for input -->
         <molecule_set_out>step2_out.xml</molecule_set_out>
         <!-- method string of calculation -->
         <method>CLUSTER_GEODIS</method>
         <options>value</optiuons>
         <!-- status:waiting|running|ready -->
         <status>waiting</status>
         <error>
            <number>0</number>
            <message></message>
         </error>
      </step>
   </job_schedule>
</job>
```

Table 1 Implemented steps

Acronym	Description
QM calculation	
AM1-GAS	AM1 gas phase optimization (MOPAC7)*
AM1-COSMO	AM1 COSMO optimization (MOPAC7)*
AM1-COSMO-SP	AM1 COSMO single point calculation (MOPAC7)*
PM3-GAS	not tested
PM3-COSMO	not tested
PM3-COSMO-SP	not tested
BP-TZVP-COSMO	BP/TZVP COSMO optimization (TM)*
BP-TZVP-GAS	BP/TZVP gas phase optimization (TM)*
BP-SVP-COSMO-SP	BP/SVP COSMO single point (TM)*
BP-SVP-GAS-SP	BP/SVP gas phase single point (TM)*
BP-SV_P-COSMO-LOOSE	BP/SV(P) cosmo optimization with quite loose conv. crit. (TM)*
BP-SV_P-GAS-LOOSE	BP/SV(P) gas optimization with quite loose conv. crit. (TM)*
BP-TZVPD-GAS-SP	BP/TZVPD single point gas phase calculation for COSMO*therm* FINE level (TM)*
BP-TZVPD-FINE-COSMO-SP	BP/TZVPD single point COSMO calculation for COSMO*therm* FINE level (TM)*
	* More information can be found in the corresponding *.def files
Conformer generation	
CF_MOPAC_CONF_GEN	CF/MOPAC7 conformer generation
BALLOON_CONF_GEN	Balloon[$] will be used for the conformer generation. The result molecule set consists of MMFF94 structures and energies.

Clustering	
CLUSTER_GEODIS	geometry clustering using the “geodis” algorithm
CLUSTER_EVNN	clustering using the energy and the nuclear-nuclear repulsion energy
CLUSTER_SMS	clustering using the sigma match similarity (COSMO results only)
CLUSTER_MU	clustering using COSMO-RS chem. potentials (COSMO results only)
Data sorting, reduction & adding	
SORT_BY_E	sort by energy
ADD_MOLECULE_SET	adds a molecule set XML (defined by the file tag, see tab 1a) to the current molecule set. The file **must** be defined. Name conflict have to be avoided by the user. The routine checks name conflicts and quits with an error if two molecules share the same name.
REDUCE_BY_E_MAX	reduce data set. Use maximal number (see definition) of molecules with a relative (to the min. conformer) energy within a defined energy window. The number of surviving molecules is defined by the tighter criterion (max number of molecules or energy window). A sort by energy will be done before the reduce algorithm starts. Therefore, the results can be expected to be sorted.
REDUCE_TO_UNIQUECODE	The unique-codes of the structures of the set are checked against the reference structure. Conformers with different uniquename than the reference structure will be neglected.
Writing	
PRINT_CONF_INFO	prints listing of molecule names and relative energies on screen (not important for calls from GUI)
WRITE_ENERGY_FILE	writes an energy file for each molecule of the current set. The structures and energies will be taken from the molecule set directly. The relative (to execution directory) path used is: `path/name.energy` with: `path`: path defined by subtag see table 1.a `name`: molecule name as defined in <molecule name =…> The level description printed to the energy files can be given in a subtag (see table 1.a). The Molecule Set `coordi-`

	`nate_file` entries will be updated
COPY_COSMO_FILE	copies the cosmo files of the relative path (to execution directory) path. `path/name.cosmo` with: `path`: path defined by subtag see table 1.a `name`: molecule name as defined in <molecule name =…> or a global name name_c0…n.cosmo (see subtag in tab. 3) if defined.
Miscellaneous	
GET_UNIQUECODE	gets the 12 character uniquecode (COSMOfrag routines) for all molecules of the set. The method will ignore errors (error numbers <0). All structures that can be read will be used. If the uniquecode calculation fails "NONAME000000" will be set instead.

[$] http://users.abo.fi/mivainio/balloon/. Mikko J. Vainio and Mark S. Johnson (2007) *Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm.* Journal of Chemical Information and Modeling, **47**, 2462 - 2474.

Table 1a: Parameters of calculation types given in table 1

Parameter tag	**Description**	**Default ***
AM1/PM3-GAS, AM1/PM3-COSMO, AM1/PM3-COSMO-SP		
`n_batch`	number of MOPAC calculations per batch (divide the multi step job into `n_batch` batches)	1000
CF_MOPAC_CONF_GEN		
`max_gas_opt`	maximum number of MOPAC gas phase calculations in first step	5000
`cf_generator_method`	defines the cf (COSMO*frag*) keywords for the conformer generation in the first step of the procedure: 0: simple method (action=3) 1: method 2 but less angles per bond rotation (rotconf=crude action=3). 2: includes rotations of important bonds (rotconf action=3) 3: method 2 but more angles per bond rotation (rotconf=fine action=3)	2

cf_enable_rotalk	enable/disable rotation of alkyle chains (0=off, 1=on)	0
n_batch	number of MOPAC calculations per batch (divide the multi step job into n_batch batches).	1000
BALLOON_CONF_GEN		
options	The base options (always used) are: verbose=0;forcefield=MMFF94.mff;fullforce=1; nInitialDimensions=6;maxtime=200000; nobadmodels=1;expand=1;contract=1; pStereoMutation=0.00 Other keywords will be added to the upper ones: **a**) via the <options> tag. E.g: <options>nconfs=90;nGenerations=99;RMSDtol=0.2</options> The options have to be separated by a semicolon. **b**) default (empty or missing option tag) a series of 7 balloon jobs will be used: 1) randomSeed=7; nconfs=100; noGA=1 2) randomSeed=1; keepInitial=1; nconfs=100; nGenerations=20; RMSDtol=0.1; pTorsionMutation=0.5; noPopulationGrowth=1 3) randomSeed=2; nconfs=100; nGenerations=100; RMSDtol=0.2; pTorsionMutation=0.2; noPopulationGrowth=1 4) randomSeed=3; nconfs=100; nGenerations=200; RMSDtol=0.3; pTorsionMutation=0.1 5) randomSeed=4; nconfs=100; nGenerations=500; RMSDtol=0.4 6) randomSeed=5; nconfs=50; nGenerations=1000; RMSDtol=0.5 7) randomSeed=6; nconfs=50; nGenerations=1000; RMSDtol=0.6 The structures of all steps will be accumulated.	see left
CLUSTER_GEODIS		
geodis_threshold1	conformers with a geodis value smaller than geodis_threshold1 will be considered as equal	0.5
geodis_threshold2	conformers with a geodis bigger smaller than	2.0

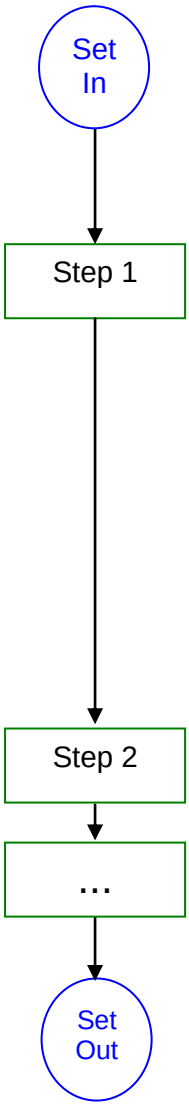
	geodis_threshold2 will be considered as different	
dihedral_threshold	conformers with a geodis value between the upper bounds will be checked by a local dihedral angle comparison. This is the max. allowed deviation in degrees.	10.0
CLUSTER_EVNN		
e_clust_thresh	energy window in kcal/mol	0.05 kcal/mol
vnn_clust_thresh	percentage of nuc.-nuc. repulsion deviation	0.05 %
CLUSTER_SMS		
sms_threshold	Sigma Match Similarity (SMS) threshold	0.95
ediel_weight	weight factor that scales the dielectric energy in the clustering procedure	1.0
CLUSTER_MU		
mu_threshold	chemical potential threshold in kcal/mol	0.2 kcal/mol
def_file	definition file name (file containing the definition of the mixtures used for the calc. of the chem. pot.). See default file for format description.	cluster_mu.def
REDUCE_BY_E_MAX		
energy_window	defines the energy window in kcal/mol	20 kcal/mol
n_max	maximal number of surviving molecules	50
REDUCE_TO_UNIQUECODE		
reference	Molecule set XML with **one** structure. The XML files needs to be located in the same directory as the input set (molecule_set_in)	no default
ADD_MOLECULE_SET		
File	defines the molecule set XML file path (relative to execution directory). This sub-tag **must** be defined.	no default
COPY_COSMO_FILE		
Path	defines the relative (to the execution dir.) path of the directory the energy files will be written to (relative to the COSMOconf execution directory). Only the last directory of the path will created automatically. An empty path (default) creates a Results_of_<job_acrnym> directory	

	(job_acronym is the name of the job definition xml file)	
global_name	The cosmo files will be sorted by energy and renamed ("global_name_cx.cosmo" (x=0,1..,n)). The "_c0" numbering will be used for single conformer compounds. An existing but empty global_name tag triggers the use of the structure set info as global name.	
WRITE_ENERGY_FILE		
path	defines the relative (to the execution dir.) path of the directory the energy files will be written to (relative to the COSMOconf execution directory). An empty path (default) creates a Results_of_<job_acrnym> directory (job_acronym is the name of the job definition xml file)	
global_name	The energy files will be sorted by energy and renamed ("global_name_cx.energy" (x=0,1..,n)). The "_c0" numbering will be used for single conformer compounds. An existing but empty global_name tag triggers the use of the structure set info as global name.	
add_comment	defines the additional info given in the 2nd line of the energy file. The string "ENERGY=number;" will be extended by the string defined in this tag. In order to be consistent with the COSMO*therm* conventions this should be: "METHOD=b-p;BASIS=def-TZVP;" for the BP-TZVP-COSMO database and "METHOD=b-p;BASIS=def-SVP;" for the BP-SVP-AM1 database	empty string
PRINT_CONF_INFO		
n_print	optional number of conf. to be printed	all conf. will be printed

* defaults defined in Job.pm

§ fixed balloon options used: --verbose=1 --forcefield=MMFF94.mff --fullforce --nInitialDimensions=6 --keepInitial --nobadmodels --randomSeed=42 --expand --contract --pStereoMutation=0.00

Example: Job definition XML




```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- general remarks:
   error number 0  -> normal termination
   error number <0 -> error, the error message should contain some
   description of the problem
-->
<job>
   <error>
      <number>0</number>
      <message></message>
   </error>
   <clean_up>1</clean_up>
   <info>first step (conf creation)</info>
   <molecule_set_in>cc_cluster_in.xml</molecule_set_in>
   <molecule_set_out>cc_cluster_out.xml</molecule_set_out>
   <job_schedule>
      <!--will be executed according to the step number-->
      <step>
      <!-- steps will be executed according to the step number -->
         <number>1</number>
         <!-- might be used in the output/error messages -->
         <info>conf. creation</info>
         <!-- file of output molecule set, not needed for input -->
         <molecule_set_out>step1_out.xml</molecule_set_out>
         <!-- method string of calculation -->
         <method>CF_MOPAC_CONF_GEN</method>
         <!-- status:waiting|running|ready -->
         <status>waiting</status>
         <error>
            <number>0</number>
            <message></message>
         </error>
      </step>
      <step>
         <number>2</number>
         <!-- just a name used in the output/error messages -->
         <info>cluster. creation</info>
         <!-- just a name used in the output/error messages -->
         <!-- file of output molecule set, not needed for input -->
         <molecule_set_out>step2_out.xml</molecule_set_out>
         <!-- method string of calculation -->
         <method>CLUSTER_GEODIS</method>
         <!-- status:waiting|running|ready -->
         <status>waiting</status>
         <error>
            <number>0</number>
            <message></message>
         </error>
      </step>
   </job_schedule>
</job>
```

Table 2: Tag description of job XML

Tag	Description
`Error`	global error description. `number` < 0 => error. The error description can be found in the `message` tag. The error on the job level contains general errors which cannot be related to the steps defined. If a specific step error occurs the job error will be set to a negative value too. => **the error definitions of the job step should be checked if the job error number <0**. An undefined error number will be interpreted as 0.
`Info`	**optional** info string
`clean_up`	reasonable clean up (1=on, 0=off) calc. directories. (**optional**, default=1)
`molecule_set_in`	input XML (see molecule set XML, In/OUT set). The relative path (to execution directory) needs to be given)
`molecule_set_out` attribute: `extractable` attribute: `directory`	output XML (see molecule set XML, In/OUT Set). The relative path (to execution directory) needs to be given). The `extractable` and `directory` attributes define the result extraction of the COSMO*conf* GUI. `extractable`: `no`: no extraction of the set `separate`: extraction to the subdirectory defined by the `directory` attribute. If the directory attribute is missing the subdirectory will be named like is the name of the set (without.xml). `join`: extraction to the general result directory (chosen by the user) `directory`: subdirectory of the general result directory that should be used if `extractable=separate` is used.
`job_schedule`	set of job steps
`Step`	definition of a job step
Subtags of job_schedule	
`Number`	The steps of the jobs will be executed according to their `number`. E.g. a step –99 will be executed before step 1, regardless of the order in the XML document.

`Info`	just some info that will be printed to the output (optional)
`molecule_set_out` attribute: `extractable` attribute: `directory`	If defined, the output structure set of this particular set will be written to the given file name. (format: molecule set XML format, the relative path (to execution directory) needs to be given) (optional). The `extractable` and `directory` attribute define the result extraction of the COSMO*conf* GUI. `extractable:` `no:` no extraction of the set `separate:` extraction to the subdirectory defined by the `directory` attribute. If the directory attribute is missing the subdirectory will be named like is the name of the set (without.xml). `join:` extraction to the general result directory (chosen by the user) `directory`: subdirectory of the general result directory that should be used if `extractable=separate` is used.
`Method`	the implemented methods are listed in table 1. The acronym from table 1 has be used here.
`Status`	This tag provides the work flow status information. Allowed values are: `waiting`, `running`, `ready`, `off`. In a new input all status values should be set to `waiting` or `off`. A missing status will be interpreted as waiting.
`Error`	job step error description. `number` < 0 => error. The error description can be found in the `message` tag. Undefined error numbers will be interpreted as 0.

Table 3: Special job requirements

Job Type (Acronym)	Special structure XML requirements
QM calculation	
AM1-GAS	-
AM1-COSMO	-
BP-TZVP-COSMO	-
BP-TZVP-GAS	-
BP-SVP-COSMO-SP	-
BP-SVP-GAS-SP	-
CF_MOPAC_CONF_GEN	only one structure
BALLOON_CONF_GEN	only one structure
Clustering	
CLUSTER_GEODIS	energy of molecule must be defined
CLUSTER_EVNN	energy of molecule must be defined
CLUSTER_SMS	only cosmo files, defined by the `coordinate_file` and `name` tag (see table 1). All cosmo/cos files must be located in the same directory.
CLUSTER_MU	only cosmo files, defined by the `coordinate_file` and `name` tag (see table 1). All cosmo/cos files must be located in the same directory.